Leukemia (2002) 16, 473–477
© 2002 Nature Publishing Group All rights reserved 0887-6924/02 $25.00
www.nature.com/leu

REVIEW

Large-scale gene expression analysis in molecular target discovery

MS Orr and U Scherf

Gene Logic Inc., Gaithersburg, MD, USA

The evolution of simple arrays consisting of a few genes to ones composed of thousands of genes and/or ESTs has allowed investigators unprecedented views of the molecular mechanisms within cells. Due to the enormous quantities of information derived from microarray analysis, new types of problems have surfaced, such as where to store all of the data. The ability to solve database or statistical problems has required the bench biologist to collaborate with database developers, software designers and statisticians to determine solutions for storage, analysis and interpretation of microarray data. The collaborative effort between these extremely diverse disciplines has led to the development of creative database query and gene expression analysis tools, producing significant reductions in the time required by researchers to filter through the datasets and discover the key processes perturbed by the diseases of interest. Both unsupervised and supervised analysis methods have been applied to gene expression data leading to the discovery of novel therapeutic targets and diagnostic markers. Furthermore, tumor classification based on their respective molecular fingerprints has led to the classification of cancer subtypes and the discovery of novel molecular taxonomies that may eventually lead to improved patient stratification and superior therapeutic strategies.
Leukemia (2002) **16,** 473–477. DOI: 10.1038/sj/leu/2402413
Keywords: microarray; hierarchical cluster analysis; principal component analysis; supervised learning; linear discriminant analysis; tissue microarray

Microarray platforms

In recent years, a variety of microarray platforms has become commercially available. The majority of pharmaceutical and biotechnology companies rely heavily on high-density chips such as cDNA arrays from Incyte, 25-mer oligonucleotide GeneChip arrays from Affymetrix, or 60-mer oligonucleotide and/or cDNA chips from Agilent technologies. The microarray chips are considered closed platforms as they can only measure genes/ESTs that have already been cloned or partially sequenced, in contrast to open systems, such as SAGE and READS that allow the discovery of unidentified genes. In the near future, complete sequence information describing every gene in the human genome will become available. Chips designed to contain every gene in the human genome, as well as other species, will be commercially available and negate the discovery advantages that open systems such as SAGE and READS currently have over microarrays.[1,2] One advantage of open systems such as READS is in their sensitivity, as open systems can currently detect low abundance genes that are below the detection limits of microarray platforms. However, the extensive labor required to use these techniques effectively inhibit their application to studies requiring high-throughput. Using open systems in a few well-defined experiments allows the identification of candidate drug targets overlooked by current microarray platforms, and they remain excellent tools for target discovery.

The cDNA chips were one of the first high-density chips to appear on the market. Incyte's LifeArray system is composed of sequences from a variety of species. The arrays are usually composed of approximately 8000–10 000 unique genes or ESTs depending on the chip selected. The technique is based on cDNA technology in which PCR products are robotically spotted on glass slides.[3] Total RNA isolated from either control or experimental conditions are fluorescently labeled with either Cy3- or Cy5-dUTP using a single round of reverse transcription. The mixture of fluorescently labeled targets is then hybridized to the glass microarray, washing steps ensue, the slide is scanned, and gene expression information is gathered and stored for analysis. Recently (October 2001), Incyte decided to terminate production of their cDNA chips. Alternative commercial sources will be required to obtain this particular type of chip in the future.

Affymetrix oligonucleotide microarrays differ from the cDNA approach in that *in situ* synthesis of 25-mer oligonucleotides is performed directly on the surface of the chip using photolithography.[4] Instead of using one cDNA PCR product to capture targets representing a particular gene of interest, multiple oligonucleotide probes are designed from various regions along the 3′ end of the gene. Furthermore, a corresponding mismatch oligonucleotide with a single base pair mismatch is designed and synthesized in a cell below the perfect match chip probe. One advantage for using oligonucleotides for chip probes is their increased specificity as compared to the larger cDNAs. The Affymetrix chips have been shown to distinguish genes with up to a 90% sequence homology allowing discrimination between the highly homologous genes that are commonly present in drug metabolism gene families.[5] Furthermore, oligonucleotides can distinguish between single nucleotide polymorphisms, as well as splice variants.[6] Similar to the cDNA sample preparation, total RNA is reversed-transcribed to cDNA. However, a major difference between the two techniques is the incorporation of the T7 promoter attached to the oligo dT primers. The T7 promoter is essential for the subsequent *in vitro* transcription reaction routinely used in the sample preparation for the Affymetrix oligonucleotide microarrays. *In vitro* transcription from the cDNA is performed in the presence of biotin-modified NTPs. The biotin-labeled cRNA produced is fragmented using a combination of heat and magnesium, and subsequently hybridized to the array. The chip is then stained with streptavidin-phycoerythrin, fluorometric scanning is performed, and the data are stored for analysis.

Recently, Rosetta Inpharmatics and Agilent Technologies have incorporated the use of ink-jet oligonucleotide synthesizer technology for microarray chip production. They are currently producing cDNA chips with approximately 14 000 chip probes, as well as an oligonucleotide chip that can contain

Correspondence: MS Orr, Gene Logic Inc., 708 Quince Orchard Road, Gaithersburg, MD 20878, USA; Fax: 301 987 1701
Received 20 September 2001; accepted 30 November 2001

SPOTLIGHT

up to approximately 49 000 unique sequences as derived from Unigene. The oligonucleotide chips use a single 60-mer probe for discriminating between genes or ESTs of interest. Two sample preparation techniques can be used. Fluorescently labeled cDNA targets or cRNA fragments have been successfully hybridized to a single 60-mer oligonucleotide and similar results were obtained with either sample preparation technique as compared to standard cDNA spotted arrays.[7] The results provide evidence that single 60-mer oligonucleotide chip probes can distinguish between the different genes or ESTs of interest.[7]

A variety of high-density microarray platforms are commercially available and all are producing enormous quantities of gene expression data. The deployment of relational databases and software interfaces that permit quick access to the rich biological information has enhanced researchers' ability to mine these reservoirs of information. Furthermore, the application of old as well as newly devised multivariate analysis tools enable researchers to quickly find the most promising candidate therapeutic targets, diagnostic markers, as well as insights into disease processes that were unimaginable a few years ago.

Unsupervised discovery techniques

Hierarchical clustering analysis (HCA)

The use of large gene expression profiling platforms has generated data in an unprecedented scale that requires alternative methods for analysis. One method that has clearly demonstrated utility for gene expression profiling experiments is the HCA technique, in association with coloring of the measured data for visualization.[8–10] The organization of massive data in a way that genes with similar expression patterns appeared next or close to each other and the representation of the underlying expression levels visualized by varying the intensity of the colored image have been extremely useful to scientists lacking experience with the analysis of multi-dimensional data sets. The individual approaches are not new as the combination of color maps representing the cluster analysis was previously used for visualization of drug–target correlations.[11] HCA alone has been previously applied to gene expression data.[12,13] Mathematically, hierarchical clustering forces data points into a strict hierarchy of nested subsets where the closest pair of points is grouped and replaced by a single point representing their set average, the next closest pair of points is treated similarly, and so on. The data points are thus fashioned into a phylogenetic tree whose branch lengths represent the degree of similarity between the sets. HCA analysis of time course expression data was initially described for *S. cerevisiae* cDNA expression data[8,9] and soon followed by data from stimulated human fibroblasts in culture.[14] Snapshots of the expression patterns contained in normal and tumor tissues profiled on oligonucleotide arrays displayed the utility of HCA in classifying genes into functional groups.[15] Alizadeh *et al*[16] and Ross *et al*[17] demonstrated the feasibility of classifying tissue according to their tissue of origin based on their respective gene expression profiles. Even gene expression patterns shared between cancer specimens and individual cell lines were found,[17] indicating the possible use of HCA for cell model system identification. HCA on gene expression data derived from human breast tumors was able to classify tumors into subtypes and provided the basis for an improved molecular taxonomy of breast cancer.[18] Potential short-comings of HCA are that there is no opportunity to re-evaluate the clustering once the process is performed and artifacts can occur due to earlier mistakes in the clustering process. It is known that the resulting trees can lock in accidental features, reflecting idiosyncrasies of the agglomeration rule. This can result in decreased robustness, non-uniqueness, and inversion problems that complicate interpretation of the hierarchy. Furthermore, the interpretation of these clusters or the recognition of the fundamental patterns, is mainly left to the observer and in-depth analysis can be a time-consuming process. Overall, the clustering approach is a powerful tool for microarray data analysis and the technique has elucidated numerous new leads in cancer research.

Principal component analysis (PCA) and multidimensional scaling (MDS)

Principal component analysis (PCA) techniques, statistical analysis tools available since the turn of the century, have been used to reduce the number of variables in multidimensional datasets, as well as for structure detection. The basic premise behind this technique is that two correlated factors are combined into one new variable. The first principal component (PC) contains the greatest variance in the dataset, the second PC describes the second most variance and so on, until one ends up with a certain number of uncorrelated principal components that can describe 100% of the variance in the system. One criticism of using this multivariate technique is that it can oversimplify patterns and subtle details may be lost. In the early phases of the discovery process, a technique that can reduce the dimensionality of a $30 \times 64\,000$ matrix down to a three-dimensional representation of the major differences in the dataset is an extremely valuable tool for microarray data mining. Furthermore, the ease and speed with which large datasets can be analyzed by this approach is a distinct advantage over other techniques, such as the hierarchical cluster analysis technique that requires larger quantities of time to process the same information. However, the two techniques can display very similar results.[19] Other multivariate techniques have a difficult time dealing with large datasets and often require enormous computing power to accomplish the analysis. A disadvantage of PCA, as a visualization tool when compared to nonlinear multidimensional scaling (MDS) technique, is that PCA has more mapping errors (distance from point to point not as accurately positioned on the plot) when placing the data in a three-dimensional plot. However, drawbacks to using the MDS technique are that the quantity of information presented in the scatter-plot is unknown and there are limitations to the size of a dataset that can be processed using a conventional computer. The linear PCA method displays the information content, as it is a linear mapping technique and millions of data points can be processed easily, which is a distinct advantage over MDS procedure. The positive and negative factors associated with any multivariate technique have to be considered before beginning the analysis procedure. However, in some cases, the ability to perform the analysis due to limitations inherent in the algorithm and/or computing power available quickly reduces the number of multivariate techniques that can be applied to the analysis of particularly large datasets.

The application of the PCA technique to the analysis of microarray data in the public domain is limited at this time. One of the first studies implementing this approach to cDNA array data found that the technique was quite useful for vis-

ualization and identification of outlier genes, when comparing expression patterns between tamoxifen-sensitive and -resistant tumors across multiple time points.[20] Researchers have applied the PCA technique to the publicly available yeast sporulation dataset, and they described similarities between groups of genes clustered by hierarchical cluster analysis and the PCA technique.[19,21] A number of recently published articles have presented the utility in using PCA in combination with other supervised learning methods such as Fischer linear discriminant analysis or artificial neural networks, techniques covered in the subsequent supervised learning section of this review.[22,23]

The MDS technique is often used for multidimensional data visualization in a three-dimensional plot based on a particular distance metric. One of the earliest applications of MDS to microarray data was utilized by Khan *et al*[24] to visualize differences between alveolar rhabdomyosarcoma cell lines. The technique has also been applied to help develop molecular classification techniques for melanoma and hereditary breast cancer.[25,26]

Supervised learning techniques

The ability to determine the best set of discriminators between clearly defined groups of samples can be accomplished using a number of different statistical techniques, such as artificial neural networks, linear discriminant analysis, gene shaving, and support vector machines (SVM), soft independent modeling of class analogy (SIMCA), or K-nearest neighbors.[27–31] Many of these techniques have been applied to biological data, but not necessarily to microarray studies.

Neural networks have been previously utilized in cancer research to predict a drug's mechanism of action based on the pattern of sensitivity across the 60 cancer cell lines utilized in the National Cancer Institute's drug screening program.[32] In gene expression studies, the combination of PCA and an artificial neural network (ANN) has been used to determine the best discriminators between subtypes of small, round, blue-cell tumors such as neuroblastoma, rhabdomyosarcoma, non-Hodgkin lymphoma, and Ewing tumors.[23] A subset of 96 genes was successfully shown to classify the samples into a particular diagnostic category.[23] The technique was also applied to the classification of estrogen receptor positive or negative breast cancer samples.[33] As many supervised learning methods overfit the models if the number of variables vastly exceeds the number of samples and this scenario is often the case with microarray data that routinely contain few samples and thousands of genes. The PCA procedure has been utilized for reducing the dimensionality of a dataset before the application of supervised learning methods, such as linear discriminant analysis or ANN. In addition to the analysis of microarray data by artificial neural networks, the combination of PCA and Fisher linear discriminant analysis techniques was shown to be capable of predicting normal and colon tumor samples correctly approximately 87% of the time.[22] Another statistical method, again based on PCA technique, is gene shaving that can be utilized in unsupervised and/or supervised fashion for rapidly identifying the best group of genes capable of distinguishing between classes of samples. Hastie *et al*[34] discovered a subgroup of genes by gene shaving that predicted the clinical survival of patients with large B cell lymphoma. Another interesting method for molecular classification of cancer is neighborhood analysis devised by Golub *et al*.[35] This technique used 'weighted votes' to determine the best discriminators between ALL and AML. The newly discovered marker genes were used to predict which class of ALL or AML samples should be placed into a particular category.[35] An assortment of supervised learning techniques can be applied to microarray data that will facilitate the development of novel techniques for classifying cancer at the molecular level and should prove useful in diagnosis, patient stratification for treatment, and patient-specific treatment approaches.

Gene expression analysis and implementation for clinical use

Numerous publications using microarray data for analysis have demonstrated the value of expression profiling to identify diagnostic markers, drug targets, or a more sophisticated way of tumor classification that has led to the development of a new field, termed 'molecular pathology'. Gene expression profiling data, in combination with drug sensitivity data, have been used in the assessment of clinical tumors for markers that may be predictors of therapeutic efficacy. The gene–drug correlation has been utilized in molecular pharmacology, to provide a rationale for selecting a particular therapeutic regimen on the basis of the molecular characteristics of a patient's tumor. This approach focuses on the sensitivity to a particular therapy rather than on the molecular consequence of therapy. Recently, two important gene–drug correlations, identified by the analysis of both a gene expression database and a drug sensitivity database, have been described 5-fluorouracil (5-FU)-dihydropyrimidine dehydrogenase (DPYD) and L-asparagine (L-ASP)-asparagine synthetase (ASNS). The antimetabolite 5-FU, commonly used to treat colorectal and breast cancer, can inhibit both RNA processing and thymidylate synthesis. Dihydropyrimidine dehydrogenase (DPYD, encoded by *DPYD*), the rate-limiting enzyme in uracil and thymidine catabolism, is also rate-limiting in 5-FU catabolism. High DPYD levels would be expected to decrease exposure of cells to the active phosphorylated forms of 5-FU. Consistent with this hypothesis, a highly significant negative correlation between *DPYD* expression and 5-FU potency has been found in the panel of 60 human cancer cell lines.[36] Closer examinations revealed that 14 of 18 cell lines with low expression of DPYD were sensitive or highly sensitive to 5-FU. Interestingly, all of the colon-derived cell lines (seven of seven) fall into that category, perhaps not coincidentally, given the clinical use of 5-FU against colon cancer. Previous DPYD enzyme activity had been assessed,[37,38] but the results using clinical materials were inconsistent.[38] The gene–drug correlation described above suggests that in the case of treatment with 5-FU, studies monitoring the expression levels of DPYD are warranted and may be a beneficial clinical marker capable of enhancing patient treatment. Another interesting gene–drug correlation in the NCI60 cell lines is the moderately high negative correlation between expression of asparagine synthetase (ASNS, encoded by *ASNS*) and L-asparaginase (L-ASP) sensitivity, with the subpanel of leukemic cell lines revealing an even more significant gene–drug correlation. The two ALL cell lines, included in the panel, expressed the lowest levels of *ASNS* mRNA and were the most sensitive to L-asparaginase. A chronic myelogenous leukemia line that had the highest expression of *ASNS* was the least sensitive to L-asparaginase. Malignant cells that lack ASNS are dependent on exogenous L-asparagine.[39] This dependence has been observed for a variety of acute lymphoblastic leukemias (ALL) and is exploited

HIGHLIGHTS

by treating ALL and other lymphoid malignancies with L-asparaginase to deplete extracellular L-asparagine.[40] Beyond the observed gene–drug correlation in the leukemia cells, there was also a slight correlation between *ASNS* expression and L-asparaginase sensitivity for the ovarian cell lines. In early clinical trials treating solid tumors like melanoma, chronic granulocytic leukemia, lymphosarcoma and reticulum cell sarcoma[40] sporadic responses to L-asparaginase have been observed. Presently in the literature, there is no evidence describing the treatment of ovarian patients with this enzyme. The L-ASP-ASNS correlation supports the possible use of ASNS expression as a marker regarding L-asparaginase therapy, especially against ovarian cancer, but perhaps also against a subset of other tumors. L-ASP, as a drug, is particularly attractive because its mechanism of action and toxicity are quite different as compared to other anticancer agents. Furthermore, newer polyethylene glycol-modified forms of the enzyme have been developed that showed improved pharmacokinetic immunologic properties than the native form used in earlier clinical trials.[41] This suggested the use of L-ASP for ovarian cancer treatment and Dr Daniel von Hoff, Arizona Cancer Center, is investigating this type of therapeutic intervention. Dr von Hoff (Arizona Cancer Center, The University of Arizona Health Sciences Center) added L-ASP to the battery of anticancer agents that are being tested against primary cultures from patient tumors (personal communication, Dr Daniel von Hoff, September 2001) and they are performing immunohistochemistry for ASNS expression in tissue arrays constructed from the same tumor.

Target validation, e-Northern analysis and future directions

A variety of traditional techniques such as Q-RT-PCR and Northern analysis are used to confirm the gene expression modulations observed by microarray analysis. Furthermore, differential expression at the protein level for the gene of interest is required as well, as alterations in message levels do not always reflect a similar level of differential expression of the protein. A classical example for a lack of agreement between message and protein levels has been extensively documented for the p53 tumor suppressor protein that acquires post-translational modifications following cellular perturbations, such as DNA damage increasing protein levels without alterations in the expression levels of the message.[42–46] Once the gene of interest is established as differentially expressed in a particular disease phenotype, the more laborious process of functionally characterizing the gene of interest by standard molecular biology methods, such as antisense or overexpression studies in the representative model system are often performed. The ability to use databases of gene expression from diverse model systems will enable researchers to rapidly identify cell lines that contain the newly discovered target gene of interest reducing the time expenditure required to screen for applicable cell systems. Furthermore, multi-tissue Northern blot is often employed to determine if a gene of interest is expressed in a particular tissue type. An '*in silico*' approach, similar to the multi-tissue Northern blot, is the e-Northern procedure that is based on the exploitation of a database of microarray gene expression information from different tissue types. This technique is routinely used at Gene Logic in the GeneExpress product. The gene of interest is queried against a panel of normal tissues, such as the heart, kidney, liver and lung etc, to determine the expression levels of the gene in various tissues. The e-Northern approach can be used as a tool to discover tissue-specific expression, a first-pass screen to prioritize candidate therapeutic targets, or a source of clues to the biological roles of ESTs. Another tool for determining if a candidate gene isolated by microarray analysis is a reasonable target for drug discovery or as a diagnostic marker is a tissue microarray. Tissue microarrays facilitate the rapid determination of the expression levels of a particular candidate gene in as many as 1000 different tissue biopsies.[47,48] This technique has the potential to increase the speed with which diagnostic or therapeutic targets are screened.[47,49,50] Even with the substantial technological advancements that have taken place, validation of microarray targets is still a major barrier blocking the full utilization of the reservoir of gene expression information from microarrays. However, an abundance of candidate drug targets and diagnostic markers are currently being discovered by microarray analysis. A steady flow of novel drug targets and diagnostic markers are being investigated and these novel experimental directions will eventually lead to improvements in cancer detection and treatment.

References

1 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995; **270**: 484–487.
2 Prashar Y, Weissman SM. Analysis of differential gene expression by display of 3′ end restriction fragments of cDNAs. *Proc Natl Acad Sci USA* 1996; **93**: 659–663.
3 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**: 467–470.
4 Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991; **251**: 767–773.
5 Gerhold D, Lu M, Xu J, Austin C, Caskey CT, Rushmore T. Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. *Physiol Genomics* 2001; **5**: 161–170.
6 Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res* 1994; **22**: 5456–5465.
7 Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001; **19**: 342–347.
8 Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998; **9**: 3273–3297.
9 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
10 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1999; **96**: 10943c.
11 Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997; **275**: 343–349.
12 Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA* 1998; **95**: 334–339.
13 Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ,

Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998; **2**: 65–73.
14 Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. The transcriptional program in the response of human fibroblasts to serum. *Science* 1999; **283**: 83–87.
15 Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999; **96**: 6745–6750.
16 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**: 503–511.
17 Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000; **24**: 227–235.
18 Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature* 2000; **406**: 747–752.
19 Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput* 2000; 455–466.
20 Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK, Fuqua SA. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J Natl Cancer Inst* 1999; **91**: 453–459.
21 Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. The transcriptional program of sporulation in budding yeast. *Science* 1998; **282**: 699–705.
22 Xiong M, Jin L, Li W, Boerwinkle E. Computational methods for gene expression-based tumor classification. *Biotechniques* 2000; **29**: 1264–1270.
23 Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001; **7**: 673–679.
24 Khan J, Simon R, Bittner M, Chen Y, Leighton S, Pohida T, Smith P, Jiang Y, Gooden G, Trent J, Meltzer P. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998; **58**: 5009–5013.
25 Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000; **406**: 536–540.
26 Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 2001; **344**: 539–548.
27 Dunn WJ III, Wold S. Structure–activity study of beta-adrenergic agents using the SIMCA method of pattern recognition. *J Med Chem* 1978; **21**: 922–930.
28 Holmes E, Nicholls AW, Lindon JC, Connor SC, Connelly JC, Haselden JN, Damment SJ, Spraul M, Neidig P, Nicholson JK. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem Res Toxicol* 2000; **13**: 471–478.
29 Unger PD, Watson CW, Liu Z, Gil J. Morphometric analysis of neoplastic renal aspirates and benign renal tissue. *Anal Quant Cytol Histol* 1993; **15**: 61–66.
30 Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, Reich M, Lander E, Mesirov J, Golub T. Molecular classification of multiple tumor types. *Bioinformatics* 2001; **17** (Suppl. 1): S316–S322.
31 Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000; **97**: 262–267.
32 Weinstein JN, Kohn KW, Grever MR, Viswanadhan VN, Rubinstein LV, Monks AP, Scudiero DA, Welch L, Koutsoukos AD, Chiausa AJ, Paull KD. Neural computing in cancer drug development: predicting mechanism of action. *Science* 1992; **258**: 447–451.
33 Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001; **61**: 5979–5984.
34 Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol* 2000; **1**: 2.
35 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–537.
36 Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000; **24**: 236–244.
37 Fischel JL, Etienne MC, Spector T, Formento P, Renee N, Milano G. Dihydropyrimidine dehydrogenase: a tumoral target for fluorouracil modulation. *Clin Cancer Res* 1995; **1**: 991–996.
38 McLeod HL, Sludden J, Murray GI, Keenan RA, Davidson AI, Park K, Koruth M, Cassidy J. Characterization of dihydropyrimidine dehydrogenase in human colorectal tumours. *Br J Cancer* 1998; **77**: 461–465.
39 Cooney DA, Handschumacher RE. L-asparaginase and L-asparagine metabolism. *Annu Rev Pharmacol* 1970; **10**: 421–440.
40 Capizzi RL, Bertino JR, Handschumacher RE. L-asparaginase. *Annual Rev Med* 1970; **21**: 433–444.
41 Wada H, Imamura I, Sako M, Katagiri S, Tarui S, Nishimura H, Inada Y. Antitumor enzyme: polyethylene glycol-modified asparaginase. *Ann NY Acad Sci* 1990; **613**: 95–108.
42 Maltzman W, Czyzyk L. UV irradiation stimulates levels of p53 cellular tumor antigen in nontransformed mouse cells. *Mol Cell Biol* 1984; **4**: 1689–1694.
43 Kastan MB, Onyekwere O, Sidransky D, Vogelstein B, Craig RW. Participation of p53 protein in the cellular response to DNA damage. *Cancer Res* 1991; **51**: 6304–6311.
44 Sakaguchi K, Herrera JE, Saito S, Miki T, Bustin M, Vassilev A, Anderson CW, Appella E. DNA damage activates p53 through a phosphorylation-acetylation cascade. *Genes Dev* 1998; **12**: 2831–2841.
45 O'Connor PM, Jackman J, Jondle D, Bhatia K, Magrath I, Kohn KW. Role of the p53 tumor suppressor gene in cell cycle arrest and radiosensitivity of Burkitt's lymphoma cell lines. *Cancer Res* 1993; **53**: 4776–4780.
46 Siliciano JD, Canman CE, Taya Y, Sakaguchi K, Appella E, Kastan MB. DNA damage induces phosphorylation of the amino terminus of p53. *Genes Dev* 1997; **11**: 3471–3481.
47 Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998; **4**: 844–847.
48 Kallioniemi OP, Wagner U, Kononen J, Sauter G. Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum Mol Genet* 2001; **10**: 657–662.
49 Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001; **412**: 822–826.
50 Horvath L, Henshall S. The application of tissue microarrays to cancer research. *Pathology* 2001; **33**: 125–129.